

DOCUMENT RESUME

ED 219 454

TM 820 483

AUTHOR Kane, Michael; Wilson, Jennifer
TITLE Errors of Measurement and Standard Setting in Mastery Testing.
PUB DATE Mar 82
NOTE 14p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, March 20-22, 1982).
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *Cutting Scores; *Error of Measurement; Mastery Tests; Scores; *Test Reliability; *Test Theory
IDENTIFIERS Domain Referenced Tests; Misclassification (Statistics); *Standard Setting

ABSTRACT

This paper evaluates the magnitude of the total error in estimates of the difference between an examinee's domain score and the cutoff score. An observed score based on a random sample of items from the domain, and an estimated cutoff score derived from a judgmental standard setting procedure are assumed. The work of Brennan and Lockwood (1980) is extended by explicitly considering the covariance between errors of measurement and errors in standard setting. In combining the two kinds of errors, it is found that the item effect included in the error of measurement is correlated with the item affect that appears in the standard setting error. Therefore, in addition to the variance components for the different sources of error, the total error variance includes the covariance between the two item effects. The implications of these results for the probability of misclassification are also discussed. (PN)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED219454

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

Errors of Measurement and Standard Setting in Mastery Testing

Michael Kane
Jennifer Wilson

Paper presented at the annual meeting of the National Council for
Measurement in Education, New York, 1982.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

M. T. Kane

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

TM 820 483

Abstract

A number of studies have estimated the dependability of domain-referenced mastery tests for a fixed cutoff score. Other studies have estimated the dependability of cutoff score ratings. Each of these two types of dependability is defined in terms of error; however, rarely are the errors for both kinds of estimation included when studying the dependability of test scores. Brennan and Lockwood's (1980) analysis of the total error in estimates of the difference between domain score and the cutoff score is extended to allow for the covariance in the two types of errors.

Errors of Measurement and Standard Setting in Mastery Testing

Introduction

Glaser and Nitko (1971) define a criterion-referenced test as one that is designed "to yield measurements that are directly interpretable in terms of specified performance standards" (p. 653). For a domain-referenced test, which is one type of criterion-referenced test, each score is interpreted in terms of the proportion of items that can be answered correctly in some domain of items. The observed score, the proportion correct on a sample of items from the domain, provides an estimate of the proportion of items in the domain that the examinee could answer correctly. Clearly, this interpretation does not depend on the performance of any reference group.

A domain-referenced test which is used to decide whether individuals have attained some particular level of performance standard is called a mastery test. The standard of performance for mastery would typically be specified by a cutoff score, defined as a proportion of the items in the domain. The decision to be made about each person is whether the person has mastered the domain of content.

It is assumed that the domain consists of a large number of discrete tasks or items and that independent random samples can be drawn from this domain. The proportion of items that the p -th person could answer correctly, if exposed to all the items in the domain, is called the domain score for the p -th person and is represented by u_p . The domain score, u_p , is a parameter defined for the person on the domain.

Mastery of the domain is defined by establishing a cutoff score, c , on the domain. A person whose domain score, u_p , is at or above the cutoff score, c , is said to be a master, and a person whose domain score is below the cutoff score is said to be a nonmaster.

Since it is generally not practical to include the entire domain in a test, the domain score is not directly observable. Rather, decisions about mastery are based on the person's performance on a sample of items from the domain. A particular random sample of items will be called a mastery test, and a person's observed score, X_{pI} , on the I -th mastery test is the proportion of items that the person answers correctly on the I -th mastery test. Although u_p is assumed to be a constant for each person, X_{pI} will typically vary from one test to another. Random samples of a particular number of items from the domain are randomly parallel tests, and the observed scores on these tests may have different means and different standard deviations over the population of persons.

In using judgemental standard setting procedures, the estimated cutoff score, Y_c , is generally based on the judgment of experts who review the items in the test and decide on an appropriate minimum passing level (MPL) for each item. The estimated cutoff score for the sample of items defining a test is simply the sum of the MPLs for the sample of items. The value of the estimated cutoff score will depend on the items that are rated and on the raters. Since the cut-off score, c , is not defined in terms of any particular group of raters but rather in terms of a wider population of qualified raters, variability among raters is counted as a source of error. Similarly, since the cutoff score is defined on the domain as a whole, variability due to the sampling of items is a source of error in estimates of the cutoff.

For the p -th person taking the I -th mastery test, a mastery decision is made by comparing the person's observed score to the estimated cutoff score, Y_c . Those persons whose observed scores are greater than or equal to Y_c are classified as masters, and those persons whose observed scores are below Y_c are classified as nonmasters. The mastery decision is therefore based on a comparison between a fallible estimate of the domain score and a fallible estimate of the cutoff score, and two kinds of misclassifications may occur. A false positive is said to occur when a nonmaster is incorrectly classified as a master (i.e., when u_p is less than c and X_{pI} is greater than or equal to Y_c). A false negative is said to occur when a master is incorrectly classified as a nonmaster (i.e., u_p is greater than or equal to c , and X_{pI} is less than c).

Almost all of the literature discussing the reliability of domain-referenced test scores assumes a fixed cutoff score, defined on the domain of content, and analyzes either the consistency of classification across tests (e.g., Hambleton and Novick, 1973; Huynh, 1976, 1978; Millman, 1973; Subkoviak, 1976), or the consistency in observed deviations from the fixed cutoff score. (e.g., Livingston, 1972; Brennan and Kane, 1977a,b; Kane and Brennan, 1980). In the first approach, the focus is on the accuracy of decisions, where accuracy is defined either in terms of the proportion of examinees who are consistently classified as masters or nonmasters, or in terms of Cohen's Kappa. The second approach takes the difference between the observed score and a constant cutoff score as an estimate of the signal to be detected, and analyzes the dependability of these estimates. Neither of these two approaches to the dependability of mastery testing includes the possibility of errors in determining the cutoff score in the analysis.

The precision of estimates of the cutoff score based on judgmental standard setting procedures has also been investigated (Andrews and Hecht, 1976; Shepard, 1980; Brennan and Lockwood, 1980). However, the errors of measurement in test scores and the errors in the standard setting procedures have generally not been addressed together. In their study of two judgmental standard setting procedures, Brennan and Lockwood (1980) considered both the errors in the test scores and the errors in the estimated cutoff score. However, they assumed that these two kinds of errors are uncorrelated. This paper evaluates the magnitude of the total error in estimates of the difference between an examinee's domain score and the cutoff score, assuming an observed score based on a random sample of items from the domain, and an estimated cutoff score derived from a judgmental standard setting procedure. The work of Brennan and Lockwood (1980) is extended by explicitly considering the covariance between errors of measurement and errors in standard setting. The implications of these results for the probability of misclassification are also discussed briefly.

Domain Scores and Errors of Measurement

The linear model given below partitions the observed scores into a number of effects. This linear model is an example of the kind of model usually employed in generalizability theory (Cronbach et.al., 1972):

$$X_{pI} = u + a_p + a_I + a_{pI} \quad (1)$$

Where u is the grand mean over persons and items, a_p is the main effect for person, p , a_I is the main effect for the sample of items, I , and a_{pI} is the interaction effect for person, p , and the item sample, I .

The use of a capital letter for the index, I, indicates a sample of n_i items rather than a single item. X_{pI} represents the person's observed score on the set of items, I. If a different set of items, J, were sampled, a different observed score, X_{pJ} , would result. If, as is usually the case, there is only one observation for each person-item combination, the $\underline{a}_{pI}$ effect includes both the person-item interaction and random error.

The model assumes that persons and items are sampled randomly and independently. Because the effects are defined in terms of deviations from mean scores, the expected value of an effect over any of its subscripts is zero, and the correlation, over persons and items, of any two of the effects is zero.

Cronbach et.al.(1972) draw a distinction between a generalizability study, or G study, and a decision study, or D study. The purpose of the G study is to estimate variance components, which can then be used to predict the dependability of measurements for a variety of D studies. The results of a G study in which persons are crossed with items would typically be stated in terms of the variance components, $\sigma^2(p)$, $\sigma^2(i)$, and $\sigma^2(pi)$, for sampling individual persons and items. For the random sample of n_i items in a D study, the variance component for persons remains the same, and the variance components for the two effects involving items are:

$$\sigma^2(I) = \sigma^2(i)/n_i \quad (2a)$$

$$\sigma^2(pI) = \sigma^2(pi)/n_i \quad (2b)$$

Note that the two components of variance in Eq(2) both decrease as the number of items increases. Since the $\underline{pI}$ effect and the I effect contribute to error, the relationship in Eq (2) is the basis for the Spearman-Brown correction; as the number of items increases, the error variance decreases and the reliability increases. In general, the equations in this paper are expressed in terms of the variance components for samples of items rather than for single items. Although this leaves the dependence of various expressions on the number of items implicit, these relationships can always be made explicit by using Eq(2).

Unbiased estimates of the variance components for the three effects, $\underline{a}_p$, $\underline{a}_I$, and $\underline{a}_{pI}$ can be obtained from a G study with persons crossed with items (all persons respond to the same set of items). If items are nested within persons (i.e., an independent sample of items is selected for each person) in the G study, the two effects $\underline{a}_I$ and $\underline{a}_{pI}$ are confounded, and their variance components cannot be estimated separately. Because of the confounding of effects that occurs in nested designs, G studies are generally most useful when they employ crossed designs, and for convenience, it is assumed that variance components are estimated from a crossed G study.

The domain score for the p-th person, u_p , is defined as the expected value of the observed score, X_{pI} , taken over the domain of items. Using the model in Eq(1), the domain score is

$$u_p = E(X_{pI}) = u + a_p \quad (7)$$

and the expected value of the domain score over the population is equal to the

grand mean, $\bar{u}$. Therefore, the domain score variance is given by:

$$E(u_p - \bar{u})^2 = E(a_p^2) \quad (8)$$

The domain score variance is the same for crossed or nested designs. The domain score variance depends on the domain definition and on the definition of the population, but it does not depend on the sampling designs used to estimate the domain scores.

Cronbach et. al. (1972, p. 84) define the error in point estimates of domain scores as the difference between the observed score and the domain score:

$$D_p = X_{pI} - u_p$$

The variance of D_p , taken over the population and the domain, is the same for nested and crossed designs and is given by:

$$o^2(D) = o^2(a_I) + o^2(a_{pI}) \quad (9)$$

Therefore, the error variance for point estimates involves both the variance of the item main effect and the person-item interaction variance.

Cronbach et.al.(1972, p. 93) define the deviation score error, d_p , as the difference between the observed deviation score and the domain deviation score:

$$d_p = (X_{pI} - \bar{u}_I) - (u_p - \bar{u}) \quad (10)$$

If items are crossed with persons, the variance in d_p is given by:

$$o^2(d) = o^2(a_{pI}) \quad (11)$$

Since the item effect is a constant for all observed scores in a crossed design, the item effect drops out of the observed deviation scores, and therefore out of the error. For a nested design, the expected observed score taken over the population is $\bar{u}$, and $o^2(d)$ is equal to $o^2(D)$ given by Eq(9); in a nested design, the average item effect is different for each observed score and thus does not cancel out in forming deviation scores.

The error, d_p , plays a prominent role in classical test theory, where interest is focused on deviation scores; in fact, $o^2(d)$ is the error variance for coefficient alpha. For mastery testing, however, where observed scores are compared to a cutoff score, which is a constant across all observations, $o^2(D)$ is a more appropriate error variance, because any error in estimating the domain score will affect this comparison. There are two sources of error in point estimates of the domain score; the main effect for items, $\bar{a}_I$, is a general systematic error affecting all observed scores in the same way, while the interaction, $\bar{a}_{pI}$ is specific to each person.

For a crossed design, the systematic error, $\bar{a}_I$, affects all observed scores in the same way, and therefore tends to bias observed scores in one direction or the other. For example, if the sample of items included in the test is particularly easy, $\bar{a}_I$ will be positive, and the observed scores will tend to be higher than they would be if the items were of average difficulty. This kind of bias will increase the probability of false positives and will decrease the probability of false negatives. It will also increase the size of the errors in estimates of the difference between domain and the cutoff score. Therefore, the systematic error, $\bar{a}_I$, is included in the error variance for mastery tests, regardless of whether the sampling of items is nested within persons or crossed with persons, and an appropriate error variance for estimates of the observed scores is $o^2(D)$.

The Cutoff Score and its Estimation

The cutoff score is typically based on the judgment of experts who review the items in the test and decide on an appropriate minimum passing level (MPL) for each item. The cutoff score for the sample of items defining a test is simply the sum of the MPLs for the sample of items. The value of the estimated cutoff score will depend on the items that are rated and on the raters. The three most commonly discussed procedures for setting standards (Ebel, 1972; Angoff, 1971; and Nedelsky, 1954) are all based on subjective judgment, and are, therefore, all influenced by variability among and within raters. An empirical study by Andrew and Hecht (1976) indicates that for a given set of items and a given group of raters, that the cutoff scores set by Ebel's procedure can be quite different from those set by Nedelsky's procedure. A study by Brennan and Lockwood (1980) indicates that Nedelsky's procedure doesn't agree with Angoff's procedure in terms of the cutoff scores that they yield.

The technology for setting the cutoff score is the weakest link in the logic of mastery testing. In the lead article of an issue of the Journal of Educational Measurement devoted entirely to setting standards, Gene Glass has discussed many of the difficulties associated with setting standards, and concludes that, "every attempt to derive a cutoff score is either blatantly arbitrary or derives from a set of arbitrary premises" (Glass, 1978, p 258). In their response to Glass's article, Popham (1978), Hambleton (1978), Block (1978), and Linn (1978) concede that standards are always arbitrary in the sense that they must be established by fallible human judgments, but they point out that standards need not be arbitrary in the sense of being capricious or unreasonable. Glass's critics make a strong case for the conclusion that reasonable cutoff scores can be established in many situations, but they do not claim that cutoff scores can be set with great precision.

The cut-off score estimate for raters, $\underline{R}$, and items, $\underline{I}$, can be represented by a linear model which parallels that given in Eq (1) for observed scores,

$$Y_{RI} = \underline{c} + \underline{c}_R + \underline{c}_I + \underline{c}_{RI} \quad (12)$$

where $\underline{c}$ is the grand mean over samples of raters and items, $\underline{c}_I$ is the main effect for items, $\underline{c}_R$ is the main effect for raters, and $\underline{c}_{RI}$ is the item-rater interaction. The "true" value of the cutoff score can be defined as the expected rating over the domain of items and over the population of raters, and is equal to the constant, $\underline{c}$. Therefore, given a domain of items and a population of raters, there are three identifiable sources of error in estimating the cutoff: the item effect, the rater effect, and the item-rater interaction. The contribution of the different sources of error is indicated by the variance in estimated cutoff scores over raters and items, which is given by:

$$\sigma^2(Y_{RI}) = \sigma^2(\underline{c}_R) + \sigma^2(\underline{c}_I) + \sigma^2(\underline{c}_{RI}) \quad (13)$$

Each of the variance components in Eq(13) represents the variance of the average value of an effect over a sample of $\underline{n}_I$ items and $\underline{n}_R$ raters. For example, the variance component for a sample of $\underline{n}_R$ raters is given by:

$$\sigma^2(\underline{c}_R) = \sigma^2(\underline{c}_r) / \underline{n}_R$$

and the variance component for a sample of n_i items is given by:

$$o^2(c_i) = o^2(c_i)/n_i$$

where $o^2(c_r)$ and $o^2(c_i)$ represent the variance components for individual raters and items respectively. The interaction variance component for samples of n_r raters and n_i items is given:

$$o^2(c_{ri}) = o^2(c_{ri})/n_i$$

where $o^2(c_{ri})$ is the interaction variance component for one rater on one item.

The Spearman-Brown correction applies to each of the three variance components taken separately, since the variance component is inversely proportional to the sample size. However, the Spearman-Brown correction does not apply to the total error variance in Eq(13), which is not inversely proportional to the number of raters, the number of items, or to the total number of observations, which is given by the product of n_i and n_r .

In this paper, it is assumed that the Angoff procedure is applied to a random sample of items from the domain, and these randomly sampled items are administered as the mastery test; this assumption forms the basis for most discussions of the reliability of mastery tests. An alternative approach that could enhance the accuracy of the decisions made with mastery tests would involve selecting items with an MPL close to 0.5. The discriminating power of an item for a population of examinees tends to be highest when the item has a difficulty level close to 0.5 for the population; items that are very easy or very difficult for a population are not very effective in making differential decisions within the population. Items with MPLs near 0.5 would tend to be maximally discriminating for the marginal candidates, because by definition, the probability that a marginal candidate will be able to answer a question is given by its MPL. By selecting items with MPLs near 0.5, the precision of the test is maximized near the cutoff score where it is most critical. Of course, if this procedure is adopted, the scores on the examination would not provide good estimates of the examinees' domain scores, since items with MPLs near 0.5 cannot be considered a random (or representative) sample from the domain. However, if the only purpose of the examination is to make mastery decisions, this would not matter.

Some authors (e.g., Taylor, Reid, Senhauser, and Shively, 1971) have suggested that the judgmental standard setting procedures (JSSP) provide a mechanism for giving any test a clear criterion-referenced interpretation. According to Glaser, a criterion-referenced test score would be interpreted in absolute terms, without reference to the performance of a reference population; by contrast, under a norm-referenced procedures, test scores are interpreted in terms of the score distribution for some population of persons. For a mastery test, a criterion-referenced interpretation would presumably indicate what it is that passing students can do and that failing students cannot do.

Judgmental standard setting procedures do not necessarily support an interpretation in terms of what examinees can do. If a judgmental standard setting procedure is applied to a test that cannot be interpreted in behavioral terms, the standard cannot be given a behavioral interpretation. In such cases, the burden of interpretation falls on a new reference

population, the population of raters, and the interpretation of the cutoff score will depend on the rules used to select raters. The JSSP's do not support an interpretation of a passing score in terms of what students can do; the results of a JSSP does not indicate the kind of behavior that distinguishes passing candidates from failing candidates, unless a behavioral interpretation of the test scores has already been developed. Although individual raters probably use some behavioral standards in setting the MPL for each item, the JSSPs do not provide a mechanism for making these behavioral standards explicit. Without an independently developed behavioral interpretation, the JSSP substitutes one reference population for another, and since the population of raters is likely to be less well defined than the reference populations of examinees typically used for non-referenced tests, the use of the JSSP's may complicate the interpretation of mastery decisions.

The Total Error

The question that is of central interest for mastery decisions is whether the difference between an examinee's domain score and the cutoff score, given by $u_p - c$, is positive or negative. The magnitude of this difference indicates the strength of the signal to be detected in making decisions about mastery. From Eqs (9) and (13), it can be seen that there are five distinct sources of error in estimating this signal. Brennan and Lockwood (1980) assumed that these five sources of error are uncorrelated. In general however, a_I , the item effect for observed scores, and c_I , the item effect for ratings of the average MPL, may be correlated; that is, the raters' estimates of how difficult an item would be for a minimally competent examinee may be related to the average difficulty of the item over the population of examinees.

The total error in estimating the difference score, $u_p - c$, is given by:

$$(X_{pI} - Y_{RI}) - (u_p - c) = a_I + a_{pI} - c_R - c_I - c_{RI} \quad (14)$$

Since the covariance between any two terms that do not have identical subscripts is zero, the total error variance is given by:

$$\sigma^2(E) = \sigma^2(a_I) + \sigma^2(a_{pI}) + \sigma^2(c_R) + \sigma^2(c_I) + \sigma^2(c_{RI}) - 2 \text{cov}(a_I, c_I) \quad (15)$$

where $\text{cov}(a_I, c_I)$ is the covariance between the item effect in observed scores and the item effect in the results of judgemental standard setting procedures.

The item effect for observed scores indicates how easy the items are, and the item effect for the ratings represents the average score expected of a hypothetical minimally competent examinee. The estimates of the probability that a minimally competent examinee answers an item correctly should be higher for items that the raters consider relatively easy and lower for items that the raters consider relatively difficult. Assuming that the raters exhibit some accuracy in evaluating the relative difficulty of items, the covariance in Eq (15) is likely to be positive, thus decreasing the total error variance.

The Effect of Errors of Measurement on the Probability of Misclassification

A detailed analysis of the effects of errors of measurement and errors in standard setting on the probability of misclassification is beyond the scope of this paper. The discussion below is intended to indicate the relevance of both types of errors to the probability of misclassification without providing the kind of detailed analysis that would require strong distributional assumptions.

Consider a person, p , whose universe score, u_p is above the cutoff score, c ; that is, $u_p - c$ is positive. What is the probability that this person will be misclassified as a nonmaster. Such a false negative result will occur for person, p , on test, I , if X_{pI} is less than Y_{RI} . We would like to approximate the probability that $X_{pI} - Y_{RI}$ is less than zero when $u_p - c$ is greater than zero.

The expected value over R and I , of the observed deviation score for person, p , is given by:

$$E_{RI} (X_{pI} - Y_{RI}) = E_{RI} X_{pI} - E_{RI} Y_{RI} = u_p - c \quad (16)$$

Therefore, the expected value of the observed deviation score is positive whenever the universe deviation score is positive. In order for a false negative to occur for person, p , the observed deviation score for person, p , must be less than the expected deviation score for person p , by an amount greater than $u_p - c$. Therefore, the probability of a misclassification for person, p , is equal to the probability of obtaining an observed deviation score which is below the mean of the observed deviation score distribution for person, p by a distance that is greater than $u_p - c$. The exact value of this probability will of course depend on the shape of the distribution of the observed deviation scores for person p ; in fact, the probability of a false negative is given by the value of the appropriate cumulative probability distribution at a point ($u_p - c$) below the mean.

However, without making any specific distributional assumptions, it is possible to draw some general conclusions about the effect of the two types of errors on the probability of misclassification. In particular, the probability of a misclassification for person, p , will be an increasing function of the variance of the person's observed deviation score distribution over samples of items and raters, which is equal to the total error variance for person, p . The total error variance in Eq (15) is the expected value over persons, of the error in estimating the difference between the universe score, u_p , and the cutoff score, c where the estimates are based on samples of n_1 items and n_r raters. Since the probability of misclassification for each person is an increasing function of the error variance for that person, the probability of misclassification for the population as a whole, should generally be an increasing function of the total error variance given by Eq(15). Therefore, except for unusual situations in which the average error variance increases while the probability of misclassification decreases, the probability of a misclassification will be an increasing function of the variances of both errors of measurement and errors in standard setting and will be a decreasing function of the covariance between these two types of errors.

Therefore, the issues discussed in this paper also relevant to the probabilities of false negatives and false positives. A more detailed analyses of the relationships between the probability of the two kinds of misclassifications and the error variances and covariances could easily be worked out for any particular of distribution.

Conclusions

This paper has raised a number of issues in criterion-referenced testing, but its main point is that the errors for criterion-referenced testing involve errors in standard setting as well as errors of measurement. In combining the two kinds of errors, it is found that the item effect, a_I , included in the error of measurement is correlated with the item affect, c_I , that appears in the standard setting error. Therefore, in addition to the variance components for the different sources of error, the total error variance includes the covariance between the two item effects, a_I and c_I .

In particular if this covariance is positive, the total error will be decreased by the magnitude of the covariance term. It would not be surprising to find a positive covariance between the difficulty level, which indicates how difficult the item is for examinees, and the MPL, which represents the expectations of how difficult the item will be for a particular subgroup of examinees, namely those who are minimally competent. Therefore the covariance term should serve to decrease the total error variance.

BIBLIOGRAPHY

- Andrews, B. J. and Hecht, J. T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 45-50.
- Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1971.
- Block, J. H. Standards and criteria: a response. Journal of Educational Measurement, 1978, 15, 291-295.
- Brennan, R. L. and Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289. (a)
- Brennan, R. L. and Kane, M. T. Signal/noise ratios for domain-referenced tests. Psychometrika. 1977, 42, 609-625. (Errata, Psychometrika, 1978, 43, 289.) (b)
- Brennan, R. L. and Lockwood, R. E. A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 1980, 4, 219-240.
- Cronbach, L. J. and Gleser, C. G. Psychological Tests and Personnel Decisions, Second Edition. Urbana: University of Illinois Press, 1965.
- Cronbach, L. J., Gleser, G. C., Nanda, M. and Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.
- Ebel, R. L. Essentials of Educational Measurement. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1972.
- Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational Measurement. Washington, D. C.: American Council on Education, 1971.
- Glass, G. V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.
- Hambleton, R. K. On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 1978, 15, 277-290.
- Hambleton, R. K. and Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.
- Huynh, H. Reliability of multiple classifications. Psychometrika, 1978, 43, 317-325.

- Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 65-78.
- Kane, M. T. and Brennan, R. L. Agreement Coefficients as Indices of Dependability for Domain-Referenced Tests. Applied Psychological Measurement, 1980, 4, 105-126.
- Linn, R. L. Demands, cautions, and suggestions for setting standards. Journal of Educational Measurement, 1978, 15, 301-308.
- Linn, R. L. Test design and analysis for measurement of educational achievement. New Directions for Testing and Measurement, 1980, 5, 81-92.
- Livingston, S. A. A criterion-referenced application of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.
- Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard setting. Review of Educational Research, 1976, 46, 133-158.
- Millman, J. Passing scores and test lengths for domain referenced measures. Review of Educational Research. 1973, 43, 205-216.
- Millman, J. and Popham, W. J. The issue of item and test variance for criterion-referenced tests: A clarification. Journal of Educational Measurement, 1974, 11, 137-138.
- Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.
- Popham, W. J. and Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.
- Popham, W. J. As always, provocative. Journal of Educational Measurement, 1978, 15, 297-300.
- Shepard, L. Standard setting issues and methods. Applied Psychological Measurement, 1980, 4, 447-467.
- Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-276.
- Taylor, D., Reid, J., Senhauser, D., Shively, J. Use of minimum pass levels on pathology examinations. Journal of Medical Education, 1971, 48, 876-881.